UNIVERSITY OF TECHNOLOGY SYDNEY,
AUSTRALIA.

DOCTORAL THESIS

Data Analytics and the Novice Programmer

Author:
Alireza AHADI

Supervisor:
Associate Professor
Raymond LISTER
(principal) and Dr Julia
PRIOR (co-supervisor)

A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy

in the

Human Centred Technology Design
School of Software

January 22, 2018

Declaration of Authorship

I, Alireza AHADI, declare that this thesis titled, "Data Analytics and the Novice Programmer" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

"When I was in college, my graduation thesis was called 'Female Directors.' I interviewed all of the important female directors from Mexico. There were four. That was it. "

Patricia Riggen

UNIVERSITY OF TECHNOLOGY SYDNEY, AUSTRALIA.

Abstract

Faculty of Engineering and Information Technology
School of Software

Doctor of Philosophy

Data Analytics and the Novice Programmer

by Alireza AHADI

The aptitude of students for learning how to program (henceforth *Programming learn-ability*) has always been of interest to the computer science education researcher. This issue of aptitude has been attacked by many researchers and as a result, different algorithms have been developed to quantify aptitude using different methods. Advances in online MOOC systems, automated grading systems, and programming environments with the capability of capturing data about how the novice programmer's behavior has resulted in a new stream of studying novice programmer, with a focus on data at large scale. This dissertation applies contemporary machine learning based analysis methods on such "big" data to investigate novice programmers, with a focus on novices at the early stages of their first semester. Throughout the thesis, I will demonstrate how machine learning techniques can be used to detect novices in need of assistance in the early stages of the semester. Based on the results presented in this dissertation, a new algorithm to profile novices coding aptitude is proposed and its' performance is investigated. My dissertation expands the range of exploration by considering the element of context. I argue that the differential patterns recognized among different population of novices is very sensitive to variations in data, context and language; hence validating the necessity of context-independent methods of analyzing the data.

CONFIRMATION OF ETHICS CLEARANCE

Human Negligible Low Risk Ethical clearance was granted for this PhD project by the University of Technology Sydney Human Research Ethics Committee under approval numbers ETH16-0340 (see Appendix D).

Acknowledgements

I would like to express my special appreciation and thanks to my adviser associate professor Raymond Lister, you have been a tremendous mentor for me. I would like to thank you for encouraging my research and for allowing me to grow as a research scientist. Your advice on my research have been priceless.

A special thanks to my family. Words cannot express how grateful I am to my father, mother and my beloved sister for all of the sacrifices that you've made on my behalf.

Alireza Ahadi

KEYWORDS

Novice Programmer
Human Factors
Measurement
Computer Science Education
Data Mining
Online assessment
MOOC
Databases
SQL Queries
Learning Edge Momentum
Bimodal Grade Distribution
Machine Learning
Programming Source Code Snapshot
Pattern Recognition
Classification
Supervised Machine Learning

Contents

List of Figures

List of Tables

List of Abbreviations

ACC	**Acc**uracy
CMS	**C**ource **M**anagement **S**ystem
CTP	**C**omputational **T**hinking **P**atterns
DM	**D**ata **M**ining
DT	**D**ecision **T**ree
EDM	Educational **D**ata **M**ining
EQ	Error **Q**uotient
FN	**F**alse **N**egative
FP	**F**alse **P**ositive
FDR	**F**alse **D**iscovery **R**atio
FNR	**F**alse **N**egative **R**atio
FPR	**F**alse **P**ositive **R**atio
IDE	**I**ntegrated **D**evelopment **E**nvironment
JAR	**J**ava **A**rchive **F**ile
LMS	**L**earning **M**anagement **S**ystem
LN	**L**ogical **N**ecessity
LS	**L**ogical **S**ufficiency
LSI	Kolb's **L**earning **S**tyle **I**nventory
ML	**M**achine **L**earning
MSLQ	**M**otivated **S**trategies **L**earning **Q**uestionnaire
NN	**N**eural **N**etwork
NPSM	**N**ormalized **P**rogramming **S**tate **M**odel
PCA	**P**rincipal **C**omponent **A**nalysis
RF	**R**andom **F**orest
RED	**R**epeated **E**rror **D**ensity
RSE	**R**osenberg **S**elf-**E**steem
SEN	**Sen**sitivity
SPC	**Spe**cificity
SQL	**S**tructured **Q**uery **L**anguage
SVM	**S**upport **V**ector **M**achine
TN	**T**rue **N**egative
TP	**T**rue **P**ositive
WATWIN	**WAT**son & God**WIN**
Web-CAT	**Web**-based **C**enter for **A**utomated **T**esting

List of Publications by Candidate

Below are the publications first authored by the Candidate which contribute to this thesis by publication.

1. Ahadi, A. and Lister, R., 2013, August. Geek genes, prior knowledge, stumbling points and learning edge momentum: parts of the one elephant?. In Proceedings of the ninth annual international ACM conference on International computing education research (pp. 123-128). ACM. (See Chapter 5)

2. Ahadi, A., Prior, J., Behbood, V. and Lister, R., 2015, June. A Quantitative Study of the Relative Difficulty for Novices of Writing Seven Different Types of SQL Queries. In Proceedings of the 2015 ACM Conference on Innovation and Technology in Computer Science Education (pp. 201-206). ACM. (See Chapter 7)

3. Ahadi, A., Lister, R., Haapala, H. and Vihavainen, A., 2015, July. Exploring machine learning methods to automatically identify students in need of assistance. In Proceedings of the eleventh annual International Conference on International Computing Education Research (pp. 121-130). ACM. (See Chapter 6)

4. Ahadi, A., Behbood, V., Vihavainen, A., Prior, J. and Lister, R., 2016, February. Students' Syntactic Mistakes in Writing Seven Different Types of SQL Queries and its Application to Predicting Students' Success. In Proceedings of the 47th ACM Technical Symposium on Computing Science Education (pp. 401-406). ACM. (See Chapter 9)

5. Ahadi, A., Prior, J., Behbood, V. and Lister, R., 2016, July. Students' Semantic Mistakes in Writing Seven Different Types of SQL Queries. In Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education (pp. 272-277). ACM. (See Chapter 8)

6. Ahadi, A., Lister, R. and Vihavainen, A., 2016, July. On the Number of Attempts Students Made on Some Online Programming Exercises During Semester and their Subsequent Performance on Final Exam Questions. In Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education (pp. 218-223). ACM. (See Chapter 11)

7. Ahadi, A., Lal, S., Leinonen, J., Lister, R. and Hellas, A. Performance and Consistency in Learning to Program. Australian Computing Education Conference, 2017 (Award winning paper for best student paper). (See Chapter 10)

8. Ahadi, A., Lister, R., Hellas, A. A Contingency Table Derived Methodology for Analyzing Course Data. 17, 3, (August 2017), 19 pages. DOI: https://doi.org/10.1145/3123814 (See Chapter 12)

THESIS BY PUBLICATION

This thesis is presented in the format of scholarly papers published during the period of my candidature, according to UTS regulations set out in the website http://uts.edu.au/current-students/dab/uts-graduate-research-school. The papers included in this thesis by publication form a research narrative which is summarized in Chapter 5. Each paper then forms a separate chapter of this thesis, inserted in its published format.

To all novice programmers. . .